The MIND Approach

Fabio Crestani

University of Strathclyde, Glasgow, UK

Open Archive Forum Workshop

Berlin, Germany, March 2003
Outline

- Project organisation
- Motivations, assumptions and main issues
- Architecture
- Searching distributed multimedia DLs with MIND
- MIND components of interest to OA Forum:
  - Resource description acquisition
  - Schema mapping
Project Organisation

- MIND: Resource Selection and Data Fusion in Multimedia Distributed Digital Libraries
- IST-RTD FP5 project
- Duration:
  - January 2001 - June 2003 (30 months)
- Project participants:
  - University of Strathclyde (UK) (coordinator)
  - University of Dortmund (Germany)
  - University of Florence (Italy)
  - University of Sheffield (UK)
  - Carnegie Mellon University (USA)
- More info at: http://www.mind-project.net/
Motivations

• Goal: enable searching hundreds of DLs using one distributed content-based access system
  – heterogeneity (in query language, schema, etc.)
  – multimedia (text, images, speech)

• Assumptions:
  – minimal level of standardisation and cooperation
    • cooperative and non-cooperative DLs
  – simplest possible query interface (Google style)
  – user interested in high precision searches
Main Issues

• Content-based access to information
• No local repository!

• Main issues:
  – resource descriptions
  – resource selection
  – schema mapping
  – data fusion
  – presentation of results and user interfaces

• Different levels of success in dealing with these issues in the project
MIND Architecture

• Design goals:
  – Distributed environment, different languages/OS
  – Modification/extension can be done easily
    • New DLs
    • New media types
    • New functionalities

• Solutions:
  – Specific components for different parts, e.g.
    • Proxy component for DL
    • Media-specific components
  – Communication via SOAP (XML, HTTP)
MIND Architecture

User Interface → Data fuser → Dispatcher → Proxy

Proxy

Proxy

Proxy

Proxy

Wrapper

Wrapper

Wrapper

Wrapper

DL

DL

DL

DL
Integration

User Interface
Data fuser
Dispatcher
Proxy
Wrapper
DL
QBS
CMU
UNIDO
USFD
DSI
USG
Resource description creation for images
Learner for schema rules parameters
Media
Media
Searching with MIND

User query

Modified user query

Prop. query
Prop. query
Prop. query
Prop. query

Query transformation

Resource selection

Database query run

Data fusion

Result list

Fused result list
Query Modification

User query → Modified user query

- Query modification
- Query transformation
- Resource selection
- Database query run
- Data fusion
Query Modification

- **Tasks:**
  - capture user information need
  - add more information about the user
  - query expansion w.r.t. relevance feedback data

- **Actions:**
  - **Interface:** captures information need in a multimedia query
  - **Dispatcher:** adds new conditions/modifies conditions and weights w.r.t. relevance feedback features
  - **Proxies:** generate DL-specific features
Query Transformation

User query → Modified user query

Query modification

Query transformation

Resource selection

Database query run

Data fusion

Prop. query

Prop. query

Prop. query

Prop. query

Prop. query

Result list

Result list

Fused result list
Why Query Transformation?

• Motivations:
  – Heterogeneous schemas
  – Thus: (uncertain) mapping between schemas to transform user query into proprietary query

• Example:
  – Dublin Core
  – RFC 1807
Query Transformation

• Task:
  – transform query w.r.t. different schemas

• Actions:
  – Proxies: transforms query condition by condition

More about this later
Resource Selection

- User query
- Modified user query
- Query modification
- Query transformation
- Resource selection
- Database query run
- Data fusion
Resource Selection

• Task:
  – find DLs that are relevant to the query
  – use decision-theoretic model:
    • use resource descriptions
    • cost factors (e.g. monetary costs, computation and communication time, retrieval quality)

• Actions:
  – **Dispatcher**: calculates for every DL the number of documents to retrieve so that overall expected costs are minimized and retrieval quality is maximised
  – **Proxies**: calculate specific costs for retrieval
Query Run

User query

Modified user query

Prop. query
Prop. query
Prop. query
Prop. query

Prop. query
Prop. query

Result list
Result list

Fused result list

Query modification
Query transformation
Resource selection
Database query run
Data fusion
Data Fusion

- User query
- Modified user query
- Query modification
- Query transformation
- Resource selection
- Database query run
- Data fusion
Data Fusion

• Task:
  – fuse together results from different DLs

• Actions:
  – Dispatcher: detects and eliminates duplicate documents (ID or content-based); modifies document weights to improve retrieval quality using global information from resource selection process
  – Data fuser: merges result lists using local information
  – Interface: creates summaries or surrogates; presents results
MIND and OAI

• Why should MIND be of some interest to OA Forum?
  – Completely different approach:
    • assumes no cooperation from DLs
    • content-based retrieval
    • no local repository of harvested metadata
  – Lessons on:
    • resource description acquisition for content-based retrieval (multimedia query-based sampling)
    • schema mapping
Creating Resource Descriptions

• Task:
  – create and update resource descriptions

• Actions:
  – Proxies:
    • start resource gathering at self-defined times
    • uses query-based sampling

• Resource descriptions are used in almost all phases of query processing
Query-base sampling

• Technique developed at CMU
• How does it work:
  1. iterative retrieval of documents using random queries
  2. assumption: union of results is representative for whole collection
  3. extract resource description w.r.t. document sample
• We have extended QBS to multimedia DLs
  – resource descriptions for images and speech
Schema mapping

• Heterogeneous DLs with different schemas require schema mapping
• MIND uses Probabilistic Datalog (UNIDO)
• Schema mapping is carried out at document and query level
  – schema mapping at document level is necessary for relevance feedback
• Handling:
  – queries/documents encoded in RDF/XML
  – transform rules into XSLT
Creating schema mapping rules

• Generating rules from the schema
  – find indicators for matching attributes
    • E.g. attribute names, datatypes (equality, sub-datatype)
  – compute probability for each attribute pair
    (Probabilistic Datalog), taking most likely candidates

• Rules created automatically, but still possible
to modifying them manually
  – significant error rate
Schema mapping at document level

Documents:

RFC 1807

- date
- author

(RFC1807,date,"01") → (DC,date,"2001")

Dublin Core
(standard schema)

- date
- creator

MARC 21

- 033
- 100
- 700
- 710
Schema mapping at query level

Queries:

Dublin Core (standard schema)
- creator/soundex

MARC 21
- 100/sounds-like
- 700/sounds-like
- 100/sounds-like

(DC,creator,soundex) \rightarrow (MARC21,100,sounds-like) with probability 0.6
Creating schema mapping rules

Standard schema

Query rules

Document rules

Built-in predicates

DL schema

Rule
11-Feb-03 → 2003-02-11
Conclusions

• MIND and OAI are very different in their assumptions about data, users, kind of searches, etc.

• Are MIND and OAI different solutions to the same problem?
  – MIND tries to deal/live with chaos
  – OAI tries to bring order in the chaos